Attorney's Docket No.: 11532-003001

APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: IMAGE METRICS IN THE STATISTICAL ANALYSIS OF DNA MICROARRAY DATA

APPLICANT: CARL S. BROWN AND PAUL C. GOODWIN

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL584781214US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

January 25, 2001
Date of Deposit

Signature

Richard E Donovan, Jr
Typed or Printed Name of Person Signing Certificate

IMAGE METRICS IN THE STATISTICAL ANALYSIS OF DNA MICROARRAY DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims benefit of U.S. Provisional Application No. 60/178,474, filed January 27, 2000.

TECHNICAL FIELD

This invention relates to DNA microarray analysis, and more particularly to using error estimates from image based metrics to analyze microarrays.

BACKGROUND

Large-scale expression profiling has emerged as a leading technology in the systematic analysis of cell physiology. Expression profiling involves the hybridization of fluorescently labeled cDNA, prepared from cellular mRNA, to microarrays carrying up to 10^5 unique sequences. Several types of microarrays have been developed, but microarrays printed using pin transfer are among the most popular. Typically, a set of target DNA samples representing different genes are prepared by PCR and transferred to a coated slide to form a 2-D array of spots with a center-to-center distance (pitch) of about 200 μm. In the budding yeast *S. cerevisiae*, for example, an array carrying about 6200 genes provides a pan-genomic profile in an area of 3 cm^2 or less. mRNA samples from experimental and control cells are copied into cDNA and labeled using different color fluors (the control is typically called green and the experiment red). Pools of labeled cDNAs are hybridized simultaneously to the microarray, and relative levels of mRNA for each gene determined by comparing red and green signal intensities. An elegant feature of this procedure is its ability to measure relative mRNA levels for many genes at once using relatively simple technology.

Computation is required to extract meaningful information from the large amounts of data generated by expression profiling. The development of bioinformatics tools and their application to the analysis of cellular pathways are topics of great interest. Several databases of transcriptional profiles are accessible on-line and proposals are pending for the development of large public repositories. However, relatively little attention has been paid

to the computation required to obtain accurate intensity information from microarrays. The issue is important however, because microarray signals are weak and biologically interesting results are usually obtained through the analysis of outliers. Pixel-by-pixel information present in microarray images can be used in the formulation of metrics that assess the accuracy with which an array has been sampled. Because measurement errors can be high in microarrays, a statistical analysis of errors combined with well-established filtering algorithms are needed to improve the reliability of databases containing information from multiple expression experiments.

DESCRIPTION OF DRAWINGS

These and other features and advantages of the invention will become more apparent upon reading the following detailed description and upon reference to the accompanying drawings.

Figure 1 is a gene expression curve according to one embodiment of the present invention.

Figure 2 is a graph illustrating the distribution of ratios calculated for each pixel of several spots according to one embodiment of the present invention.

Figure 3 is a graph illustrating the sport intensity standard deviation plotted against the average signal.

Figure 4 is a graph plotting covariance as a function of variance according to one embodiment of the invention.

DETAILED DESCRIPTION

The present invention involves the process of extracting quantitatively accurate ratios from pairs of images and the process of determining the confidence at which the ratios were properly obtained. One common application of this methodology is analyzing cDNA Expression Arrays (microarrays). In these experiments, different cDNA's are arrayed onto a substrate and that array of probes is used to test biological samples for the presence of specific species of mRNA messages through hybridization. In the most common implementation, both an experimental sample and a control sample are hybridized simultaneously onto the same probe array. In this way, the biochemical process is controlled for throughout the experiment. The ratio of the experimental hybridization to the control

hybridization becomes a strong predictor of induction or repression of gene expression within the biological sample.

The gene expression ratio model is essentially,

$$Expression\_Ratio = \frac{Experiment\_Expression}{Wild-type\_Expression}$$

Equation 1

In typical fluorescent microarray experiments, levels of expression are measured from the fluorescence intensity of fluorescently labeled experiment and wild-type DNA. A number of assumptions are made about the fluorescence intensity, including: 1) the amount of DNA bound to a given spot is proportional to the expression level of the given gene; 2) the fluorescence intensity is proportional to the concentration of fluorescent molecules; and 3) the detection system responds linearly to fluorescence.

By convention, the fluorescent intensity of the experiment is called "red" and the wild-type is called "green". The simplest form of the gene expression ratio is

$$R = \frac{r}{g}$$

Equation 2

where r and g represent the number of experiment and control DNA molecules that bind to the spot. R is the expression ratio of the experiment and control.

In real situations, however, r and g are unavailable. The measured values, r_m and g_m, include an unknown amount of background intensity that consists of background fluorescence, excitation leak, and detector bias. That is,

$$r_m = r + r_b$$

Equation 3

$$g_m = g + g_b$$

Equation 4

where r_b and g_b are unknown amounts of background intensity in the red and green channels, respectively.

Including background values, the gene expression ratio becomes:

$$R = \frac{r_m - r_b}{g_m - g_b}$$

Equation 5

Equation 5 shows that solving the correct ratio R requires knowledge of the background intensity for each channel. The importance of the determining the correct background values is especially significant when r_m and g_m are only slightly above than r_b and g_b. For example, the graph 100 in Figure 1 shows the effect of a ten count error in determining r_b or g_b, in the case where R is known to be one. The expected result is shown as line 105. A ten count error in the denominator is shown as line 110. A ten count error in the numerator is shown as 115.

In experimental situations, the sensitivity of the gene expression ratio technique can be limited by background subtraction errors, rather than the sensitivity of the detection system. Accurate determination of r_b and g_b is thus a key part of measuring the ratio of weakly expressed genes.

Rearranging equation 5, gives

$$r_m = R(g_m - g_b) + r_b = Rg_m + k$$

Equation 6

where

$$k = r_b - Rg_b .$$

Equation 7

Least squares curve-fit of equation 6 can be used to obtain the best-fit values of R and k, assuming that r_b and g_b are constant for all spot intensities involved with the curve-fit. The validity of this assumption depends upon the chemistry of the microarray. Other background intensity subtraction techniques, however, can have more severe limitations. For example, the local background intensity is often a poor estimate of a spot's background intensity.

Two approaches have been taken to the selection of spots involved with the background curve-fit. Since most microarray experiments contain thousands of spots, of which only a very small percent are affected by the experiment, it is probably best to curve-fit all spots in the microarray to equation 6. A refinement of this method is to use all spots that have no process control defects. Another alternative is to include ratio control spots within the array and use only those for curve fitting. The former two are preferred, because

curve fitting either the entire array or at least much of it yields a strong statistical measurement of the background values. In the case where the experiment affects a large fraction of spots, however, it may be necessary to use ratio control spots.

Constant k is interesting because it consists of a linear combination of all three desired values. While it is not possible to determine unique values of r_b and g_b from the curve-fit, there are two types of constraints that can be used to select useful values. First, the background values must be greater than the bias level of the detection system and less than the minimum values of the measured data. That is,

Constraint 1:

$$r_{bias} \le r_b \le r_{m_{\min}}$$

Equation 8

$$g_{bias} \le g_b \le g_{m_{\min}}$$

Equation 9

The second type of constraint is based on the gene expression model. For genes that are unaffected by the experiment and are near zero expression, both the experiment and the control expression level should reach zero simultaneously. In mathematical terms, when $r \rightarrow 0$, then $g \rightarrow 0$. A linear regression of $(r_m - r_b)$ versus $(g_m - g_b)$ should then yield a zero intercept. That is, selection of appropriate r_b, g_b should yield linear regression of

$$(r_m - r_b) = m(g_m - g_b) + b$$

Equation 10

such that b is approximately zero. This occurs when

$$b = mg_b - r_b$$

Equation 11

The pair of values r_b and g_b that create a zero intercept of the linear regression is thus the second constraint that can be used for extracting the background subtraction constants. Solving equations 7 and 11 for g_b gives

Constraint 2:

$$g_b = \frac{k+b}{m-R}$$

Equation 12

where R and k come from the best-fit of 6, and m and b come from linear regression of 10. The background level r_b can then be calculated by inserting g_b into equation 7 or 11.

Although it is almost always possible to generate a curve-fit of the microarray spot intensities, it is not always possible to satisfy constraints 1) and 2), especially at the same time. Failure to satisfy constraint 1) is an indication that the experiment does not fit the expected ratio model or that one of the linearity assumptions is untrue.

A somewhat trivial explanation of a failure to satisfy the constraints is that the spot intensities have been incorrectly determined. A common way that this happens is that the spot locations are incorrectly determined during the course of analysis.

Under ideal circumstances, one would also expect that the linear regression slope, m, should equal the best-fit ratio R. This can also be used as a measure of success. At the same time, the linear regression intercept b should equal zero when the r_b and g_b meet constraint 1).

Ratio Distribution Statistics

The measured values of the numerator and denominator are random variables with mean and variance. That is,

$$r_m - r_b = \bar{r} \pm r_{SD}$$

$$g_m - g_b = \bar{g} \pm g_{SD}$$

Equation 13

where r and g are mean values and r_{SD} and g_{SD} are standard deviations of r and g. The ratio R of r and g is then a random variable too with an expected value R_E and variance R_{SD}. That is,

$$\mathrm{R} = R_E \pm R_{SD} = \frac{\bar{r} \pm r_{SD}}{\bar{g} \pm g_{SD}}$$

Equation 14

Assuming that the measurement of numerator and denominator are normally distributed variables, an estimate of R_E and R_{SD} can be formed from Taylor series expansion.

$$R_E \approx \frac{\bar{r}}{\bar{g}} + g_{SD}^2 \frac{\bar{r}}{\bar{g}^3} - \frac{\sigma_{rg}}{\bar{g}^2}$$

Equation 15

$$R_{SD} \approx \sqrt{g_{SD}^2 \frac{\bar{r}^2}{\bar{g}^4} + \frac{r_{SD}^2}{\bar{g}^2} - 2\sigma_{rg} \frac{\bar{r}}{\bar{g}^3}}$$

Equation 16

where σ_{rg} is the covariance of the numerator and denominator summed over all image pixels in the spot.

$$\sigma_{rg} = \frac{1}{n} \sum_{i=1}^{n} (r_i - \bar{r})(g_i - \bar{g})$$

Equation 17

Coefficient of Variation

Assessing the quality of microarray scans and individual spots within an array is an important part of scanning and analyzing arrays. A useful metric for this purpose is the coefficient of variation (*CV*) of the ratio distribution, which is simply

$$CV = \frac{R_{SD}}{R}$$

Equation 18

In effect, the *CV* represents the experimental resolution of the gene expression ratio. Minimizing the *CV* should be the goal of scanning and analyzing gene expression ratio experiments. Minimization of R_{SD} is the best way to improve gene expression resolution. The graph 200 in Figure 2 gives two examples of comparisons between scanned images. The ratios of a first spot are shown in line 205, while the ratios of a second spot are shown as line 210. The first spot has a CV of 0.21, while the second spot has a CV of 0.60. The distribution of the ratios in the first spot 205 have a narrower distribution than the rations of the second spot 210. Thus, the first spot is a better spot to analyze.

Equation 16 shows that the variability of the ratio decreases dramatically as a function of *g*, which is a well understood phenomenon. Dividing by a noisy measurement that is near zero produces a very noisy result.

The ratio variance has an interesting dependence on the covariance σ_{rg}. Large values of σ_{rg} reduce the variability of the ratio. This dependence on the covariance is not widely known. In the case of microarray images, strong covariance of the numerator and denominator is a result of three properties of the image data: good alignment of the

numerator and denominator images, genuine patterns and textures in the spot images, and a good signal-to-noise ratio (r/r_{SD} and g/g_{SD}).

Table 1 summarizes how variables combine to reduce R_{SD}.

Variable	R_{SD} Reduction
g	↑
r	↓
g_{SD}	↓
r_{SD}	↓
σ_{rg}	↑

Table 1: A short summary of the direction that variables need to move in order to reduce the ratio distribution variance.

Spot CV

The *CV* is a fundamental metric and represents the spread of the ratio distribution relative to the magnitude of the ratio. Figure 3 shows that even though the second spot has a higher ratio than the first spot, the CV is 3X higher. The uncertainty of the second spot's ratio is far greater than the first spot's ratio. Thus, in addition to being useful for separating spots from the control population, the *CV* can also serve as an independent measure of a spot's quality.

Average CV

The average CV of the entire array of spots gives an excellent metric of the entire array quality. Scans from arrayWoRx alpha systems have been shown to have approximately ¼ the average CV of a corresponding laser scan.

Normalized Covariance

Covariance is known to be an indicator of the registration among channels, as well as the noise. Large covariance is normally a good sign. Low covariance, however, doesn't always mean the data are bad; it may mean that the spot is smooth and has only a small amount of intensity variance. Likewise, high variance is not necessarily bad if the variance is caused by a genuine intensity pattern within the spot. Figure 3 demonstrates that standard deviation increases with increasing spot intensity; the dependence is approximately linear. Figure 3 also shows that the observed standard deviation is not simply caused by the statistical noise associated with counting discrete events (statistical noise). Spots that have a substantial intensity pattern caused by non-uniform distribution of fluorescence will have a

large variance and a large covariance (if the detection system is well aligned and has low noise).

Thus, to make the covariance and the variance values useful they must be normalized somehow. In general, this can be accomplished by dividing the covariance by some measure of the spot's intensity variance. To determine the spot's variance, one could select one of the channels as the reference (for example the control channel, which is green), or one could use a combination of the variance from all channels. The following table gives examples of the normalized covariance calculation:

Normalization Method	Normalized covariance calculations	
Variances added in quadrature	$\sigma'_{rg} = \frac{\sigma_{rg}}{\sqrt{\sigma_r^2 + \sigma_g^2}}$	Equation 19
Variances added	$\sigma'_{rg} = \frac{\sigma_{rg}}{\left[\frac{(\sigma_r + \sigma_g)}{2}\right]}$	Equation 20
Control channel variance only	$\sigma'_{rg} = \frac{\sigma_{rg}}{\sigma_g}$	Equation 21
Experiment channel variance only	$\sigma'_{rg} = \frac{\sigma_{rg}}{\sigma_r}$	Equation 22

where, σ'_{rg} is the normalized covariance, and σ_r, and σ_g are the variances of channels 1 and 2, respectively.

Figure 4 illustrates a plot 400 of all the spot's covariance values versus their average variance (as in equation 20). The plot 400 of Figure 4 reveals that the normalized covariance is a very consistent value. The slope of the points in Figure 4 gives the typical value of the normalized covariance. (The average of the normalized covariance would give a similar result.) Outliers on the graph are almost always below the cluster of points along the line. Such outlying points occur when the intensity variance of the spot is unusually high, relative to the covariance. A study of these points shows that they have some sort of defect, which is often a bright speck of contaminating fluorescence.

Covariance/Variance Correlation of the Entire Array

Systematically poor correlation between covariance and variance can also point to the scanner's inability to measure covariance due to poor resolution, noise, and/or channel misalignment. Linear regression of the points in Figure 4 gives an indication of the scanner's ability to measure covariance. A broad scatter plot obviously indicates poor correlation: the variance of the spot intensities is inconsistent between the channels. A low slope indicates that the scanner has relatively high variance, relative to it's ability to measure covariance. Thus, one could compare scanners by comparing the slope and correlation coefficient of a linear regression of Figure 4 (when the same slide is scanned). A good scanner has a tight distribution with large slope and outliers indicate array fabrication quality problems rather than measurement difficulties. The average and standard deviation of the normalized covariance give similar results and could be used instead of the slope and correlation coefficient, respectively.

Spot Intensity Close to Local Background

Spots that are close, or equal, to local background may be indistinguishable from background. A statistical method is employed to determine whether pixels within the spot are statistically different than the background population.

Spot Intensity Below Local Background

Spot intensities below the local background are a good example of how the local backgrounds are not additive. Such spots are not necessarily bad, but are certainly more difficult to quantify. This is a case where proper background determination methods are essential. The method described above can make use of such spots, provided that there is indeed signal above the true calculated background.

Ratio Inconsistency (Alignment Problem or "Dye Separation")

This metric compares the standard method of measuring the intensity ratio with an alternative method. The standard method uses the ratio of the average intensities, as described above. The alternative measure of ratio is the average and standard deviation of the pixel-by-pixel ratio of the spot. For reasonable quality spots, these ratios and their respective standard deviations are similar.

There are two main source causes of inconsistency. Either the slide preparation contains artifacts that affect the ratio, or the measurement system is unable to adequately

measure the spot's intensity. The following table lists more details about each source of inconsistency.

Slide Preparation	**Measurement Problems**
Probe separation	Noise
Target separation	Misregistration
Contamination with fluorescent material	Non-linear response between channels.

Note that all the problems listed in the table will also reduce the amount of covariance. In the case of slide preparation problems, the ratio inconsistency points to chemistry problems, whereas measurement problems point to scanner inadequacy.

Numerous variations and modifications of the invention will become readily apparent to those skilled in the art. Accordingly, the invention may be embodied in other specific forms without departing from its spirit or essential characteristics.